Towards Universal Fake Image Detectors that Generalize Across Generative Models

Year: 2023 | Citations: 391 | Authors: Utkarsh Ojha, Yuheng Li, Yong Jae Lee

---

Abstract

With generative models proliferating at a rapid rate, there is a growing need for general purpose fake image detectors. In this work, we first show that the existing paradigm, which consists of training a deep network for real-vs-fake classification, fails to detect fake images from newer breeds of generative models when trained to detect GAN fake images. Upon analysis, we find that the resulting classifier is asymmetrically tuned to detect patterns that make an image fake. The real class becomes a ‘sink’ class holding anything that is not fake, including generated images from models not accessible during training. Building upon this discovery, we propose to perform real-vs-fake classification without learning; i.e., using a feature space not explicitly trained to distinguish real from fake images. We use nearest neighbor and linear probing as instantiations of this idea. When given access to the feature space of a large pretrained vision-language model, the very simple baseline of nearest neighbor classification has surprisingly good generalization ability in detecting fake images from a wide variety of generative models; e.g., it improves upon the SoTA [50] by +15.07 mAP and +25.90% acc when tested on unseen diffusion and autoregressive models. Our code, models, and data can be found at https://github.com/Yuheng-Li/UniversalFakeDetect